4-19-2019

Simulation-Based Inference on Mixture Experiments

Tejasv Bedi
txb3485@rit.edu

ROCHESTER INSTITUTE OF TECHNOLOGY

MASTERS THESIS

Simulation-Based Inference on Mixture Experiments

Author:
Tejasv BEDI

Supervisor:
Dr. Robert PARODY

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science

in

Applied Statistics
College of Science
Department of Mathematical Sciences

April 19, 2019

ROCHESTER INSTITUTE OF TECHNOLOGY

Abstract

Dr. Robert Parody
School of Mathematical Sciences

Masters of Science

Simulation-Based Inference on Mixture Experiments

by Tejasv BEDI

Mixture Experiments provide a foundation to optimize the predicted response based on blends of different components . Parody and Edwards (2006) gave a method of inference on the expected response of a 2nd-order rotatable design, utilizing a simulation-based critical point to give substantially sharper intervals when compared to the simultaneous confidence intervals provided by Sa and Edwards (1993). Here, we begin with discussing the theory of mixture experiments and pseudocomponents. Then we move on to review the literature of simulation-based methods for generating critical points and visualization techniques of general response surface designs. Next, we develop the simulation-based technique for a $\{q, 2\}$ Simplex-Lattice Design and visualize the simulation-based confidence intervals for the expected improvement in response based on two examples. Finally, we compare the efficiency of the simulation-based critical points relative to Scheffé's adaptation of critical points for the general response surface. We conclude by providing an efficiency table and demonstrate superiority of the simulation-based method over the Scheffé's adaptation on the basis of sample size savings.

Acknowledgements

First and foremost, I would like to thank my research advisor Dr. Robert Parody who motivated this research idea and guided me in completing my work towards the thesis. My thesis committee members, Dr. Ernest Fokoué and Dr. Carol Marchetti have also been there to get the best out of me and to make me a better researcher. I would like to thank all the professors and staff of the Statistics family for making a memorable learning experience for me at RIT. Finally, I would thank my friends and family for showering their love and blessings on me and providing me with emotional support that I always required throughout this journey.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Mixture Experiments

Over the years, experimentation in fields such as analytical chemistry, industrial engineering and, applied mathematics and statistics has revolved around optimizing some operational factors or components to obtain desirable properties in the final product, under some given experimental conditions. Response Surface Experiments provide a foundation to extract meaningful relationships between several explanatory variables and one or more response variables. Once such an experiment is designed, it allows a practitioner to discover the desirable settings of the explanatory variables that optimize a given set of response variables. To mathematically model this experiment, Box and Wilson suggested a second-degree polynomial model as an approximation to such experiments.

Mixture Experiments are a special case of response surface experiments that allows an experimenter to optimize the proportions of ingredients for a fixed quantity of those ingredients. The response variable in these experiments only depend on the proportions of the ingredients and not their total amount being used. These experiments are modeled as polynomials with the given restriction that all the component proportions must add up to one. Given the nature of the model and the restriction applied to it, suppose we are including q experimental components to formulate a product, then the experimental region with $q = 3$ components can be plotted as a triangle, for $q = 4$ we will obtain a tetrahedron, and so on.

Let us consider an example of a Mixed fruit juice and the ingredients used to create one. Suppose we combine three fruits namely apple, lemon and orange to create a Mixed Fruit drink. It would be of the greatest interest to the manufacturer to discover the perfect blend that creates a flavor, well received by the consumers. The flavor of a drink is highly sensitive to the proportion of components it comprises of. Therefore, different manufacturers advertise similar fruit drinks that are quite different in taste because of a different composition of ingredients. Mixture Experiments help manufactures in optimizing desired response variables based on the mixture of component blends of substances which is generally hard to capture . Getting back to the example, suppose, the manufacturer wants to market the Mixed Fruit drink as having a highly tangy flavor. By running a mixture experiment, he will be able to discover if any particular ingredient is the root cause of tanginess in the drink, or if any blend of certain ingredients is the reason for that flavor which the manufacturer is looking for. There is a huge possibility that lemon or orange alone might not contribute much towards the desirable flavor, but it could be possible that a binary blend of both the ingredients in equal proportions $(50\% : 50\%)$ are extracting the maximum amount of desirable flavor. To capture such blending properties of

ingredients that optimize the final product, mixture experiments become a necessity. The manufacturer would have missed on such an important result without the application of such experiments.

1.2 Simulation Experiments

In recent times, technology is improving at an exponential rate, trying to meet the requirements of the industry and the working sector. There has been a lot of focus on increasing processing power and speed to enable multitasking and carrying out computationally heavy operations that were once only possible in theory. One such domain was that of Simulation Experiments. Traditionally, research in such a field was only restricted to theoretical aspects of statistics. Having to apply these techniques on data sets and extracting meaningful results involved a lot of time and cost. This was one of the biggest drawback of these methods. At present, all computers are being incorporated with an amazing processing power, that too at a modest price. This is because it has become an absolute necessity for running modern day applications, and carrying out multitasking and efficient programming. For instance, a social networking site could be using complex algorithms, neural networks and artificial intelligence in ways that one could never imagine. Hence, Simulation Methods that were previously not considered feasible, demonstrate a lot of research potential and modern day applications to statistical problems. Monte Carlo Simulation methods help in solving complex problems using random sampling, optimization, numerical procedures and probability distributions. Simulation is generally applied in situations when a closed form solution is unobtainable using the usual mathematical tools. We would be using this technique for our research method to encounter the same problem.

1.3 Study Layout

The idea behind this study is to establish the use of a simulation technique to the realm of Mixture Designs, being a modern topic of research in the field of Experimental Design. The next chapter discusses the theory behind Mixture Models and the designs associated with them. Chapter three is geared towards providing a comprehensive review of existing methods and techniques that are used for this study. Chapter four explains the theory behind the methodology for this topic of research and its application to datasets. The final chapter provides the final results and conclusion to this research, and discusses the shortcomings, along with the scope of expanding on this topic for future research.

Chapter 2

Background

2.1 Introduction to Mixture Experiments and Models

The concept of mixture experiments came into existence when there was a need to mathematically formalize experiments that involved mixing various ingredients or components, to gain insight on the properties of each blend individually, as well as their various combinations with each other. To approach these designs from a modeling perspective, we associate them with polynomial models that were introduced by Scheffé in the early years (1958-1965). This approach has been believed to be the foundation of mixture experiments, as claimed by many renowned scientists and research scholars. The most essential property of these models that differentiate them from other polynomial models is the addition of a specific restriction on the input space. The restriction being that, all the proportions of component blends must add up to unity, i.e. $x_1 + x_2 + \ldots + x_q = 1$. Using this constraint, we are able to modify the polynomial model equations, and observe some interesting properties of the mixture model. We would now proceed with discussing a basic type of a mixture design.

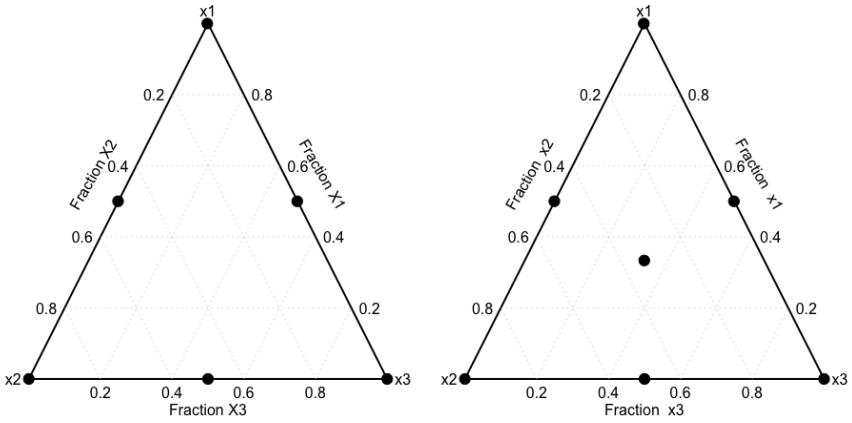
2.1.1 Simplex-Lattice Design

A simplex-lattice design is used when a polynomial equation is used to represent a response surface over an entire simplex region. The points plotted on the region are equally spaced and have the restriction of all the components adding up to 1. A $\{q, m\}$ simplex-lattice design can be expressed as a design with q components and $m + 1$ equally spaced points such that the component proportions are

$$x_i = 0, \frac{1}{m}, \frac{2}{m}, \ldots, 1$$

where m can be defined as the highest degree of blending included in the design space. For eg. for binary blending we have $m = 2$, for ternary blending we have $m = 3$ and so on. To visualize a $\{q, m\}$ simplex-lattice design, we illustrate two examples in Figure 2.1. In this figure, the design points for pure blends (x_1, x_2, x_3) are given by $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$ such that 1 denotes the presence of that particular component in the order of (x_1, x_2, x_3) and 0 denotes the absence of the components in that particular blend. For pure blends, only one component is considered at a time. Considering binary blends, two components are blended together in equal proportions at a time. Hence, the components (x_1, x_2, x_3) will take values $(0.5, 0.5, 0)$, $(0.5, 0, 0.5)$ and $(0, 0.5, 0.5)$ suggesting binary blending of components (x_1, x_2), (x_1, x_3) and (x_2, x_3) respectively. In Figure (A), the vertices of the triangle represents the pure blends, while the 3 mid-points are the binary blends between the 3 components. In Figure (B), we add a centroid in the design space given by

$(1/3, 1/3, 1/3)$. The centroid represents the ternary blending of the all the 3 components in equal proportions.



(A) {3,2} Simplex-Lattice Design (B) {3,3} Simplex-Lattice Design

FIGURE 2.1: Experimental regions of a $\{3, m\}$ Simplex-Lattice Designs

2.1.2 Simplex-Lattice Model Equations

The general equation of an m^{th} degree polynomial regression model is given by

$$Y(\mathbf{x}) = \beta_0 + \sum_{i=1}^{q} \beta_i x_i + \sum_{i \leq j}^{q} \sum^{q} \beta_{ij} x_i x_j + \sum_{i \leq j \leq k}^{q} \sum^{q} \sum^{q} \beta_{ijk} x_i x_j x_k + \ldots \tag{2.1}$$

$Y(\mathbf{x})$ is the response, given a vector of known values $\mathbf{x}$; βs are the population parameters or the model coefficients that are fixed but unknown and $x_i, x_j, x_k, \ldots$ are explanatory variables that are known to us.

The key here is to derive the equation for a $\{q, m\}$ simplex-lattice design by multiplying some of the terms of equation (2.1) by the restriction $(x_1 + x_2 + \ldots + x_q) = 1$. The resulting equation would be addressed as the "canonical" form of the polynomial equation. To demonstrate the proof, we will consider linear and quadratic regression models and use the restrictions to derive model equations for $\{q, 1\}$ and $\{q, 2\}$ simplex-lattice designs respectively.

Considering a linear regression model

$$Y(\mathbf{x}) = \beta_0 + \sum_{i=1}^{q} \beta_i x_i \tag{2.2}$$

The model restriction is given by

$$\sum_{i=1}^{q} x_i = 1 \tag{2.3}$$

Substituting (2.3) in (2.2) we have

$$Y(\mathbf{x}) = \beta_0 \left(\sum_{i=1}^{q} x_i \right) + \sum_{i=1}^{q} \beta_i x_i \tag{2.4}$$

$$= \sum_{i=1}^{q} (\beta_0 + \beta_i) x_i$$

$$Y(\mathbf{x}) = \sum_{i=1}^{q} \beta_i^* x_i \tag{2.5}$$

Where $\beta_i^* = \beta_0 + \beta_i$ for all $i = 1, 2, \ldots q$

Now we work on a quadratic regression equation to derive a $\{q, 2\}$ simplex-lattice, in a similar fashion.

The second degree polynomial equation is given by

$$Y(\mathbf{x}) = \beta_0 + \sum_{i=1}^{q} \beta_i x_i + \sum_{i=1}^{q} \beta_{ii} x_i^2 + \sum_{i<j}^{q} \sum \beta_{ij} x_i x_j \tag{2.6}$$

Now, modifying equation (2.3) we have

$$x_i + \sum_{j \neq i}^{q} x_j = 1$$

$$x_i = 1 - \sum_{j \neq i}^{q} x_j \tag{2.7}$$

$$x_i^2 = x_i \left(1 - \sum_{j \neq i}^{q} x_j \right) \tag{2.8}$$

Substituting equations (2.3), (2.7) and (2.8) in (2.6), we get

$$Y(\mathbf{x}) = \beta_0 \left(\sum_{i=1}^{q} x_i \right) + \sum_{i=1}^{q} \beta_i x_i + \sum_{i=1}^{q} \beta_{ii} x_i \left(1 - \sum_{j \neq i}^{q} x_j \right) + \sum_{i<j}^{q} \sum \beta_{ij} x_i x_j$$

$$= \sum_{i=1}^{q} (\beta_0 + \beta_i + \beta_{ii}) x_i - \sum_{i=1}^{q} \beta_{ii} x_i \left(\sum_{j \neq i}^{q} x_j \right) + \sum_{i<j}^{q} \sum \beta_{ij} x_i x_j$$

$$Y(\mathbf{x}) = \sum_{i=1}^{q} \beta_i^* x_i + \sum_{i<j}^{q} \sum \beta_{ij}^* x_i x_j \tag{2.9}$$

Using a similar derivation technique we can also obtain the equation of $\{q, 3\}$ cubic model given by

$$Y(\mathbf{x}) = \sum_{i=1}^{q} \beta_i^* x_i + \sum_{i<j}^{q} \sum \beta_{ij}^* x_i x_j + \sum_{i<j}^{q} \sum \delta_{ij} x_i x_j (x_i - x_j) + \sum_{i<j<k}^{q} \sum \beta_{ijk}^* x_i x_j x_k \tag{2.10}$$

$$Y(\mathbf{x}) = \sum_{i=1}^{q} \beta_i^* x_i + \sum_{i<j}^{q} \sum \beta_{ij}^* x_i x_j + \sum_{i<j<k}^{q} \sum \beta_{ijk}^* x_i x_j x_k \tag{2.11}$$

where, (2.11) is a special case of a cubic model in which the term $\delta_{ij}x_ix_j(x_i - x_j)$ is not considered.

From this point onward, we will remove the asterisks (*) from β coefficients, as they were just used to differentiate the simplex model equations from general polynomial equations.

Now we explain the significance of the coefficients of the simplex-lattice model equations. We first consider the simple case of $\{q, 1\}$ and $\{q, 2\}$ models. Suppose, we have pure components without the presence of blending. Then, in equations (2.5) or (2.9), if we consider a component i, we will substitute $x_i = 1$, this would result in $x_j = 0$ for all $j \neq i$. Hence, we obtain $Y(\mathbf{x}) = \beta_i$. Therefore, β_i can be defined as the expected change or response to the pure component i. Moving on to a situation of linear blending. Suppose there exists a linear blending between components i and j, then the model equation is represented by $Y(\mathbf{x}) = \beta_i x_i + \beta_j x_j$ (using (2.5)), where x_i and x_j add up to 1 and $x_k = 0$ for all $k \neq i, j$. There could be a situation that by using equation (2.5), the model is under fitting the data, which could result in a loss of information. The reason of this situation might be the unaccounted presence of two-way interactions or binary blending. Then, by fitting equation (2.9) instead, we will obtain $Y(\mathbf{x}) = \beta_i x_i + \beta_j x_j + \beta_{ij} x_i x_j$. The term $\beta_{ij} x_i x_j$ could be computed by taking the difference between equations (2.9) and (2.5). If the excess, represented by the term $\beta_{ij} x_i x_j$ is positive, or $\beta_{ij} > 0$, then the excess is considered as the synergism of the binary mixture, where β_{ij} is the second-order model coefficient of binary synergism. On the contrary, if $\beta_{ij} < 0$, the deficit is called the antagonism of the binary mixture. Similarly, if a cubic model is better suited for the situation, then in equation (2.10), the term $\delta_{ij}x_ix_j(x_i - x_j)$ represents an excess or synergism. While, δ_{ij} is the cubic coefficient of the binary synergism between x_i and x_j. If $\delta_{ij} \neq 0$, the term $\delta_{ij}x_ix_j(x_i - x_j)$ could take negative as well as positive values resulting in synergistic and antagonistic blending between the two components. The term $\beta_{ijk}x_ix_jx_k$ represents ternary blending in the model, where β_{ijk} is the third order coefficient of ternary synergism.

In the next section, we will discuss about the parameter estimation of simplex-lattice design models.

2.2 Parameter Estimation for a $\{q, 2\}$ Simplex-Lattice Design

The parameters in the $\{q, m\}$ polynomials are expressible as simple functions of the expected responses at the points of the $\{q, m\}$ simplex-lattice designs. In this section, we will discuss the parameter estimation for a $\{q, 2\}$ Simplex-Lattice Design that involves a quadratic model equation.

To obtain the model estimates, we would use the method of least squares (OLS). This procedure involves in computing the residuals using the model equation and then summing up the square of the residual terms. The final step involves optimizing the squared term w.r.t to the model parameters and solving the equations to obtain the OLS estimates. The procedure could be demonstrated mathematically, as follows

The model equation for a quadratic model can be written as

$$y_u = \sum_{i}^{q} \beta_i x_i + \sum_{i<j}^{q}\sum^{q} \beta_{ij} x_i x_j + \varepsilon_u \tag{2.12}$$

Let the equation of predicted response be given as

$$\widehat{y}_u = \sum_{i=1}^{q} \hat{\beta}_i x_i + \sum_{i<j}^{q}\sum^{q} \hat{\beta}_{ij} x_i x_j \tag{2.13}$$

Where $\hat{\beta}_i$ and $\hat{\beta}_{ij}$ are the estimates of β_i and β_{ij} respectively.

Let the residuals be denoted as e_u, for all $u = 1, 2, \ldots, r_i$; where r_i is the total number of replications of the i^{th} blend. Then,

$$\begin{aligned} e_u &= y_u - \widehat{y}_u \\ e_u^2 &= (y_u - \widehat{y}_u)^2 \\ \sum_{u=1}^{r_i} e_u^2 &= \sum_{u=1}^{n} (y_u - \widehat{y}_u)^2 \end{aligned}$$

According to the OLS principle, we have,

$$\begin{aligned} \widehat{\boldsymbol{\beta}}^{OLS} &= \underset{\hat{\beta}_i, \hat{\beta}_{ij}}{\arg\min} \left\{ \sum_{u=1}^{r_i} e_u^2 \right\} \\ &= \underset{\hat{\beta}_i, \hat{\beta}_{ij}}{\arg\min} \left\{ \sum_{u=1}^{r_i} (y_u - \widehat{y}_u)^2 \right\} \\ &= \underset{\hat{\beta}_i, \hat{\beta}_{ij}}{\arg\min} \left\{ \sum_{u=1}^{r_i} \left(y_u - \sum_{i=1}^{q} \hat{\beta}_i x_i - \sum_{i<j}^{q}\sum^{q} \hat{\beta}_{ij} x_i x_j \right)^2 \right\} \end{aligned} \tag{2.14}$$

Here, $\widehat{\boldsymbol{\beta}}^{OLS}$ is the vector of OLS estimates for all βs.

The optimization of (2.14) becomes much simpler when we apply the restriction given by equation (2.3). Now, let $E = \sum_{u=1}^{r_i} e_u^2$ for simplicity. The optimization goes as follows

Finding the partial derivative of E w.r.t $\hat{\beta}_i$ using (2.14), & equating it to 0

$$\begin{aligned} \frac{\partial E}{\partial \hat{\beta}_i} &= -2x_i \sum_{u=1}^{r_i} \left(y_u - \sum_{i=1}^{q} \hat{\beta}_i x_i - \sum_{i<j}^{q}\sum^{q} \hat{\beta}_{ij} x_i x_j \right) \\ &= -2x_i \sum_{u=1}^{r_i} \left(y_u - \hat{\beta}_i x_i - \sum_{i' \neq i, i'=1}^{q} \hat{\beta}_{i'} x_{i'} - \sum_{i<j}^{q}\sum^{q} \hat{\beta}_{ij} x_i x_j \right) \end{aligned}$$

Applying restriction (2.3) for a pure blend i.e. if $x_i = 1$, then, $x_{i'}, x_j = 0$, for all $(i', j) = 1, 2, \ldots, q$ such that $i' \neq i$ and $i < j$

$$\frac{\partial E}{\partial \hat{\beta}_i} = -2 \sum_{u=1}^{r_i} (y_u - \hat{\beta}_i) = 0$$
$$b_i = \hat{\beta}_i^{OLS} = \frac{\sum_{u=1}^{r_i} y_u}{r_i} = \bar{y}_i \tag{2.15}$$

Where $b_i = \bar{y}_i$ is the OLS estimate of β_i, for all $i = 1, 2, \ldots, q$.

Now, we would find the OLS estimates for binary blends when we have, $(x_i, x_j) = 1/2$ and $x_k = 0$.

Finding the partial derivative of E w.r.t $\hat{\beta}_{ij}$ using (2.14), & equating it to 0

$$\begin{aligned}
\frac{\partial E}{\partial \hat{\beta}_{ij}} &= -2x_i x_j \sum_{u=1}^{r_{ij}} \left(y_u - \sum_{i=1}^{q} \hat{\beta}_i x_i - \sum_{i<j}^{q} \sum \hat{\beta}_{ij} x_i x_j \right) \\
&= -\frac{1}{2} \sum_{u=1}^{r_{ij}} \left(y_u - \frac{\hat{\beta}_i}{2} - \frac{\hat{\beta}_j}{2} - \frac{\hat{\beta}_{ij}}{4} \right) = 0 \\
&= \sum_{u=1}^{r_{ij}} y_u - \frac{r_{ij}}{2} (\hat{\beta}_i + \hat{\beta}_j) - \frac{r_{ij} \hat{\beta}_{ij}}{4} = 0 \\
\hat{\beta}_{ij} &= 4 \left(\frac{\sum_{u=1}^{r_{ij}} y_u}{r_{ij}} \right) - 2(\hat{\beta}_i + \hat{\beta}_j) \\
&= 4\bar{y}_{ij} - 2(\hat{\beta}_i + \hat{\beta}_j)
\end{aligned} \tag{2.16}$$

Using (2.15), we would substitute the OLS estimates of β_i and β_j, in (2.16)

$$b_{ij} = \hat{\beta}_{ij}^{OLS} = 4\bar{y}_{ij} - 2(\bar{y}_i + \bar{y}_j) \tag{2.17}$$

Hence, $b_{ij} = \hat{\beta}_{ij}^{OLS} = 4\bar{y}_{ij} - 2(\bar{y}_i + \bar{y}_j)$ is the OLS estimate of β_{ij}, for all $(i, j) = 1, 2, \ldots, q$ such that $i < j$.

2.3 Moments of Parameter Estimates

The properties of the moments of the least squares estimates in (2.15) and (2.17) depend on the distributional properties of the random errors ϵ_u. We have assumed that the errors ϵ_u, for all u, are uncorrelated and identically distributed with mean zero and variance σ^2 i.e. $\epsilon_u \sim N(0, \sigma^2)$. Thus, the mean, variance and covariance of the estimates b_i and b_{ij} are derived as follows

2.3.1 Expectation of Parameter Estimates

The expectation of b_i is derived using (2.12), (2.15) and by applying the assumption $\mathbb{E}(\epsilon_u) = 0$

$$\begin{aligned}
\mathbb{E}(b_i) &= \mathbb{E}(\bar{y}_i) \\
&= \mathbb{E}\left(\frac{\sum_{u=1}^{r_i} y_u}{r_i}\right) \\
&= \frac{\sum_{u=1}^{r_i} \mathbb{E}(y_u)}{r_i} \\
&= \frac{1}{r_i}\sum_{u=1}^{r_i}\left(\sum_{i=1}^{q}\beta_i x_i + \sum_{i<j}^{q}\sum^{q}\beta_{ij}x_i x_j\right) \qquad (2.18)
\end{aligned}$$

Applying restriction (2.3) for a pure blend to (2.18) i.e. if $x_i = 1$, then, $x_{i'}, x_j = 0$, for all $(i', j) = 1, 2, \ldots, q$ such that $i' \neq i$ and $i < j$

$$\begin{aligned}
\mathbb{E}(b_i) &= \frac{1}{r_i}\sum_{u=1}^{r_i}\beta_i \\
&= \frac{1}{r_i}(r_i\beta_i) \\
\mathbb{E}(b_i) &= \beta_i \qquad (2.19)
\end{aligned}$$

From (2.19), we follow that the OLS estimator b_i is an unbiased estimator of β_i.

Now, the expectation of b_{ij} is derived using (2.12), (2.17), (2.19) and by applying the assumption $\mathbb{E}(\epsilon_u) = 0$

$$\begin{aligned}
\mathbb{E}(b_{ij}) &= \mathbb{E}(4\bar{y}_{ij} - 2(\bar{y}_i + \bar{y}_j)) \\
&= 4\left(\frac{\sum_{u=1}^{r_{ij}}\mathbb{E}(y_u)}{r_{ij}}\right) - 2(\beta_i + \beta_j) \\
&= \frac{4}{r_{ij}}\sum_{u=1}^{r_{ij}}\left(\sum_{i=1}^{q}\beta_i x_i + \sum_{i<j}^{q}\sum^{q}\beta_{ij}x_i x_j\right) - 2(\beta_i + \beta_j) \qquad (2.20)
\end{aligned}$$

Applying restriction (2.3) for a binary blend to (2.20) i.e. if $(x_i, x_j) = 1/2$, then $x_k = 0$.

$$\begin{aligned}
\mathbb{E}(b_{ij}) &= \frac{4}{r_{ij}}\sum_{u=1}^{r_{ij}}\left(\frac{(\beta_i + \beta_j)}{2} + \frac{\beta_{ij}}{4}\right) - 2(\beta_i + \beta_j) \\
&= \frac{4}{r_{ij}}(\frac{r_{ij}\beta_{ij}}{4}) \\
\mathbb{E}(b_{ij}) &= \beta_{ij} \qquad (2.21)
\end{aligned}$$

From (2.21), we follow that the OLS estimator b_{ij} is an unbiased estimator of β_{ij}.

2.3.2 Variance of Parameter Estimates

The variance of b_i is derived using (2.12), (2.15) and by applying the assumption $\mathbb{V}(\epsilon_u) = \sigma^2$

$$\begin{aligned}
\mathbb{V}(b_i) &= \mathbb{V}(\bar{y}_i) \\
&= \mathbb{V}\left(\frac{\sum_{u=1}^{r_i} y_u}{r_i}\right) \\
&= \frac{1}{r_i^2}\sum_{u=1}^{r_i}\mathbb{V}(y_u) \\
&= \frac{1}{r_i^2}\sum_{u=1}^{r_i}\mathbb{V}(\epsilon_u) \\
&= \frac{1}{r_i^2}(r_i\sigma^2) \\
\mathbb{V}(b_i) &= \frac{\sigma^2}{r_i} \qquad (2.22)
\end{aligned}$$

Using (2.17), (2.22) and following a similar procedure as above, we can also obtain

$$\begin{aligned}
\mathbb{V}(b_{ij}) &= \mathbb{V}(4\bar{y}_{ij} - 2\bar{y}_i - 2\bar{y}_j) \\
&= \frac{16\sigma^2}{r_{ij}} + \frac{4\sigma^2}{r_i} + \frac{4\sigma^2}{r_j} \qquad (2.23)
\end{aligned}$$

For the case of equal replications for each blend, we have,

$$\mathbb{V}(b_{ij}) = \frac{24\sigma^2}{r} \qquad (2.24)$$

2.3.3 Covariance of Parameter Estimates

We can estimate the covariance between the parameter estimates for a pair of pure blends; a pair consisting of a pure blend and a binary blend; and a pair of binary blends. The results are obtained as follows

$$\mathbb{COV}(b_i, b_j) = \mathbb{COV}(b_i, b_{jk}) = \mathbb{COV}(b_{ij}, b_{kl}) = 0. \qquad (2.25)$$

The covariance between coefficient estimates with different subscripts is 0 because there is no dependency between them.

For a pair consisting of a pure blend and a binary blend having one subscript in common, we have,

$$\begin{aligned}
\mathbb{COV}(b_i, b_{ij}) &= \mathbb{E}[\bar{y}_i(4\bar{y}_{ij} - 2\bar{y}_i - 2\bar{y}_j)] - \mathbb{E}(\bar{y}_i)\mathbb{E}(4\bar{y}_{ij} - 2\bar{y}_i - 2\bar{y}_j) \\
&= -2\mathbb{E}(\bar{y}_i^2) + 2(\mathbb{E}(\bar{y}_i))^2 \\
&= -2\mathbb{V}(\bar{y}_i) \\
&= \frac{-2\sigma^2}{r_i} \qquad (2.26)
\end{aligned}$$

Similarly, For a pair consisting of binary blends having one subscript in common, we have,

$$\mathbb{COV}(b_{ij}, b_{jk}) = \frac{4\sigma^2}{r_i} \tag{2.27}$$

2.4 Estimate of Predicted Response $\widehat{Y}(\mathbf{x})$

In this section we discuss about deriving an expression for the variance of the predicted response for given values of $\mathbf{x}$. As the estimates of model parameters are random variables, the predicted response $\widehat{Y}(\mathbf{x})$ is also random. To obtain a simplified form of $\widehat{Y}(\mathbf{x})$, we replace the parameters with their estimated values. We would further notice that computing the variance is much easier after simplifying the model equation.

The estimate of response is given by

$$\begin{aligned}
\widehat{Y}(\mathbf{x}) &= \sum_{i=1}^{q} b_i x_i + \sum_{i<j}^{q}\sum^{q} b_{ij} x_i x_j \\
&= \sum_{i=1}^{q} \bar{y}_i x_i + \sum_{i<j}^{q}\sum^{q} (4\bar{y}_{ij} - 2\bar{y}_i - 2\bar{y}_j) x_i x_j \\
&= \sum_{i=1}^{q} \bar{y}_i \left[x_i - 2x_i \left(\sum_{j\neq i}^{q} x_j \right) \right] + \sum_{i<j}^{q}\sum^{q} 4\bar{y}_{ij} x_i x_j \\
&= \sum_{i=1}^{q} a_i \bar{y}_i + \sum_{i<j}^{q}\sum^{q} a_{ij} \bar{y}_{ij}
\end{aligned} \tag{2.28}$$

Where $a_i = x_i(2x_i - 1)$ and $a_{ij} = 4x_i x_j$ for all $i, j = 1, 2, \ldots, q, i < j$. The terms a_i and a_{ij} are fixed as they only depend on $\mathbf{x} = (x_1, x_2, \ldots, x_q)'$. Since $\bar{y}_i$ and $\bar{y}_{ij}$ are averages of r_i and r_{ij} observations (replicates) respectively. By making substitutions in (2.12), variance of $\widehat{Y}(\mathbf{x})$ can be written as

$$\mathbb{V}[\widehat{Y}(\mathbf{x})] = \sigma^2 \left\{ \sum_{i=1}^{q} \frac{a_i^2}{r_i} + \sum_{i<j}^{q}\sum^{q} \frac{a_{ij}^2}{r_{ij}} \right\} \tag{2.29}$$

Where $\mathbb{V}(\bar{y}_i) = \sigma^2/r_i$ and $\mathbb{V}(\bar{y}_{ij}) = \sigma^2/r_{ij}$. If we have equal number of replications for all the blends, equation (2.29) can be simplified even further.

$$\mathbb{V}[\widehat{Y}(\mathbf{x})] = \frac{\sigma^2}{r} \left\{ \sum_{i=1}^{q} a_i^2 + \sum_{i<j}^{q}\sum^{q} a_{ij}^2 \right\} \tag{2.30}$$

2.5 Pseudocomponents

The concept of L-pseudocomponents arises from the idea of restricting the simplex region with a smaller simplex within that region itself. This concept can be applied by restricting at least one of the components with a lower bound greater than 0 i.e. $0 \leq L_i \leq x_i$ for all $i = 1, 2, \ldots, q$. The range of the L-pseudocomponents could be defined by $r_L = 1 - L$, where $L = \sum_{i=1}^{q} L_i$. Then the following transformation is

used to obtain L-pseudocomponents from the full simplex region:-

$$x_i' = \frac{(x_i - L_i)}{1 - L} = \frac{(x_i - L_i)}{r_L} \text{ for } i = 1, ..., q \text{ and L<1} \tag{2.31}$$

As it is easier to demonstrate and interpret a 2-dimensional simplex, we would consider a $\{3, m\}$ Simplex-Lattice Design that is restricted by the lower bounds $x_1 \geq 0.4$, $x_2 \geq 0.2$ and $x_3 \geq 0.2$. Lawson and Willden (2016) provide us with a graphical package to visualize mixture designs. Figure 1 is obtained using the package 'mixexp' in R provided by the mentioned authors. The L-pseudocomponent is demonstrated by the smaller triangle within the original simplex. If we were to apply the transformation on a higher dimensional simplex, we will obtain a smaller simplex contained inside the full simplex region, having the same dimensionality.

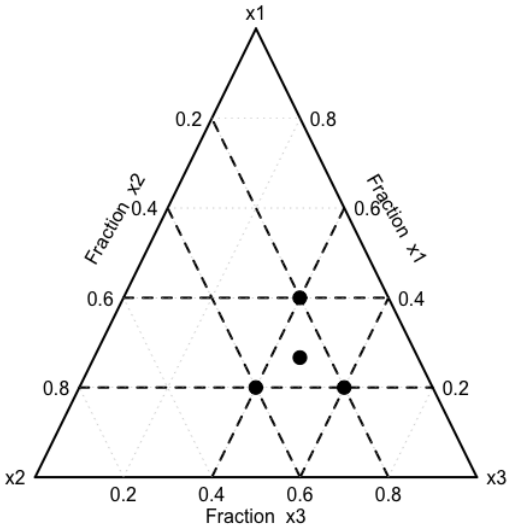


FIGURE 2.2: The simplex region based on the constraints $x_1 \geq 0.2$, $x_2 \geq 0.2$ and $x_3 \geq 0.4$

If the design region is bounded by one or more upper bounds, U-pseudocomponents are used. This concept can be applied by restricting at least one of the components with an upper bound less than 1 i.e. $x_i \leq U_i \leq 1$ for all $i = 1, 2, \ldots, q$. The range of the U-pseudocomponents could be defined by $r_U = U - 1$, where $U = \sum_{i=1}^{q} U_i$. Then the following transformation is used to obtain U-pseudocomponents from the full simplex region:-

$$x_i' = \frac{(U_i - x_i)}{U - 1} = \frac{(U_i - x_i)}{r_U} \tag{2.32}$$

for all $i = 1, 2, \ldots, q$ and $U > 1$.

We notice that the orientation of the resulting experimental region is the reverse of the original mixture space. At times, the new experimental region won't be completely contained by the original mixture space. Hence, the points inside the experimental region will not fall inside the original simplex as the model restriction (2.3) is not met. To check if such a situation occurs, we will see if

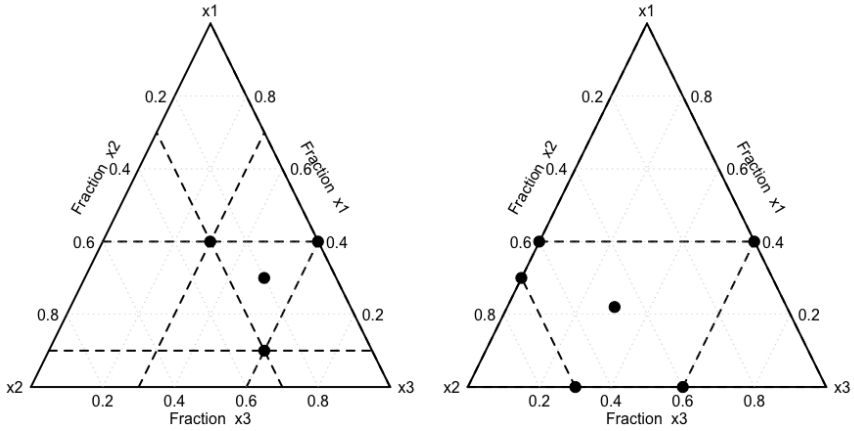
$$U - U_{min} \leq 1 \tag{2.33}$$

where U_{min} is the smallest of all the upper bounds. If (2.33) is not met, we would remove the points that fall outside the original simplex. If (2.33) is met, then we wouldn't need to remove points as all of them will be contained inside the original simplex.

Table 1 provides two examples for the two different cases mentioned above. In Example 1, as restriction (2.33) is met, we will notice that the smaller triangle is inside the original simplex. While, in Example 2, the smaller triangle isn't contained inside the original simplex, as the restriction is not met. Figure 2(a) illustrates the restricted simplex for Example 1, and Figure 2(b) illustrates the restricted simplex for Example 2.

TABLE 2.1: Upper Bound examples

Example	Bounds	$U - U_{min}$
1	$x_1 \leq 0.6, x_2 \leq 0.3, x_3 \leq 0.4$	1.0
2	$x_1 \leq 0.6, x_2 \leq 0.7, x_3 \leq 0.4$	1.3



(A) (B)

FIGURE 2.3: Experimental regions for the upper bound examples, (A) experimental region for example 1; (B) experimental region for example 2

If we compare the figures of U-pseudocomponents with that of L-pseudocomponents, we would notice that the smaller triangles are flipped in orientation, in the case of upper restrictions.

If we were to consider both upper and lower restrictions, then the resulting region would be the intersection of the two individual regions. To obtain such a region, we would first include the region based on lower restrictions, as in such a case, all the points will always fall inside the original simplex. Then we would find points based on upper restrictions, and include only those points that are inside the original simplex. Finally, we would take the intersection between the region of points obtained using the lower restrictions and the upper restrictions.

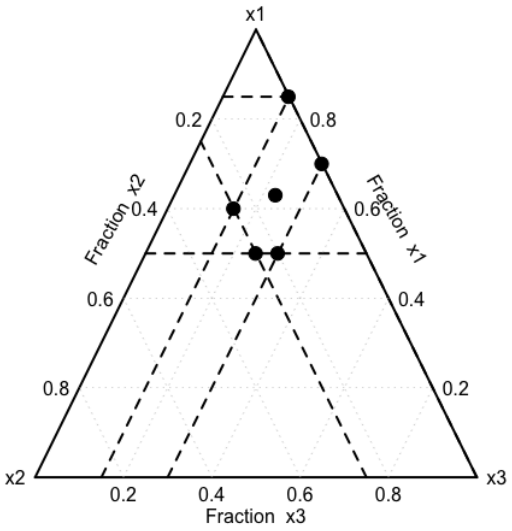


FIGURE 2.4: Experimental region for the example with both lower and upper bounds, restricted by $0.15 \leq x_1 \leq 0.3, 0 \leq x_2 \leq 0.25$ and $0.5 \leq x_3 \leq 0.85$

Now we demonstrate an example assuming that the components are restricted both ways by $0.15 \leq x_1 \leq 0.3, 0 \leq x_2 \leq 0.25$ and $0.5 \leq x_3 \leq 0.85$. The plot in Figure 2.3 was obtained by removing the points that do not fall into the intersection of the simplexes. We notice that the shape of the region in Figure 2.3 is not similar to the shape of the original simplex. We used examples with 3 components as it was easy to visualize them on a two-dimensional scale. Box and Draper (2007); Cornell (2002) provide a deeper explanation on pseudocomponents and the theory of mixture experiments.

2.6 Cholesky's Decomposition

Cholesky's Decomposition is a method of factorizing a matrix into a product of two triangular matrices. This techniques is widely used for simplifying matrix inversion and cutting down on the run time of computer programs that involve inverting matrices.

In order to implement this technique, the matrix under consideration must be Hermitian and positive-definite i.e. a necessary and sufficient condition for a complex matrix $\mathbf{A}$, to be positive definite is that the Hermitian part

$$\mathbf{A}_H \equiv \frac{1}{2}(\mathbf{A} + \mathbf{A}^H)$$

The Cholesky decomposition of a Hermitian positive-definite matrix $\mathbf{A}$ can be obtained by the form $\mathbf{A} = \mathbf{G}\mathbf{G}'$, where $\mathbf{G}$ is a lower triangular matrix with real and positive diagonal entries, and $\mathbf{G}'$ is the conjugate transpose of $\mathbf{G}$. Every real-valued symmetric positive-definite matrix or every Hermitian positive-definite matrix has a unique Cholesky decomposition.

Following is the element-wise decomposition of the matrix equation $\mathbf{A} = \mathbf{G}\mathbf{G}'$, using Cholesky's Decomposition

$$\begin{bmatrix} A_{00} & A_{01} & A_{02} \\ A_{10} & A_{11} & A_{12} \\ A_{20} & A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} G_{00} & 0 & 0 \\ G_{10} & G_{11} & 0 \\ G_{20} & G_{21} & G_{22} \end{bmatrix} \begin{bmatrix} G_{00} & G_{01} & G_{02} \\ 0 & G_{11} & G_{12} \\ 0 & 0 & G_{22} \end{bmatrix}$$

where,

$$G_{jj} = \sqrt{A_{jj} - \sum_{k=0}^{j-1}(G_{jk}^2)}$$
$$G_{ij} = \frac{1}{G_{jj}}\left(A_{jj} - \sum_{k=0}^{j-1} G_{ik}G_{jk}\right)$$

This technique will be further discussed in Chapters 3 and 4.

Chapter 3

Literature Review

3.1 Simulation-Based Multiple Comparisons

This section discusses how the idea for constructing simulation-based critical points was introduced by Edwards and Berry (1987). They proclaimed that the renowned methods for creating simultaneous confidence intervals provided very conservative solutions in general. Hence, they laid down the foundation for simulation-based multiple-comparisons for One-Way Analysis of Variance and Analysis of Covariance models.

To explain this further, Edwards and Berry (1987) defined a vector of unknown model parameters $\boldsymbol{\beta}' = (\beta_1, \beta_2, \ldots, \beta_k)$ and their estimates $\hat{\boldsymbol{\beta}}' = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k)$ having a multivariate normal distribution with mean $\boldsymbol{\beta}$ and covariance matrix $\sigma^2\mathbf{V}$, where $\mathbf{V}$ is known. They defined the estimate of variance $\hat{\sigma^2}$ to be independent of $\hat{\boldsymbol{\beta}}$, such that $\nu\hat{\sigma^2}/\sigma^2$ has a χ^2_ν distribution. Hence, they defined the natural pivotal quantity for a linear combination $\theta_j = \mathbf{c}_j'\boldsymbol{\beta}$ at $(1-\alpha)\times 100\%$ confidence level. Where, $\mathbf{c}_j = (c_{j1}, c_{j2}, \ldots, c_{jk})$ is a vector of contrasts for all $j = 1, 2, \ldots, p$.

$$W = \max_{1\leq j\leq p}\left\{\frac{|\mathbf{c}_j'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})|}{\hat{\sigma}\sqrt{\mathbf{c}_j'\mathbf{V}\mathbf{c}_j}}\right\} \tag{3.1}$$

Then in theory, it must be possible to compute the upper-α percentile point, w_α, satisfying $P(W > w_\alpha) = \alpha$. Hence, the exact intervals for θ_j for all $j = 1, 2, \ldots, p$ are given by

$$\mathbf{c}_j'\hat{\boldsymbol{\beta}} \pm w_\alpha\sqrt{\mathbf{c}_j'\mathbf{V}\mathbf{c}_j} \tag{3.2}$$

Although, it was realized that the exact solution for w_j can not be easily determined numerically or analytically. Therefore, methods that could provide conservative approximations for w_j were being used instead. Therefore, Edwards and Berry (1987) suggested to substitute a random variable W_α instead of the exact pivotal quantity w_α. To obtain W_α the following Lemma was defined.

Lemma 1. Let $W_1, W_2, , W_m, W$ be independent random variables, each with the same continuous probability distribution. For specified α, let $r = (m+1)(1-\alpha)$, take α, m such that r is an integer. If $W_{(1)} \leq W_{(2)} \leq \ldots, \leq W_{(m)}$ are the order statistics of $W_1, W_2, \ldots, W_m$, then $P(W > W_\alpha) = \alpha$.

The Lemma suggests that if we simulate m iterations of the random variable W, where m is large enough. Then by substituting $W_\alpha = W_{(r)}$ instead of w_α in equation (2.16), we have exact confidence level $(1-\alpha)\times 100\%$. Then the random variable W

is defined by the following two equations

$$\mathbf{u}_j = \frac{\mathbf{G}'\mathbf{c}_j}{\sqrt{\mathbf{c}_j'\mathbf{V}\mathbf{c}_j}} \qquad \text{for all } j = 1, 2, \ldots, p \tag{3.3}$$

$$W = \max_j \left\{ \frac{\mathbf{u}_j'\mathbf{Z}}{Y} \right\} \tag{3.4}$$

where $\mathbf{Z}$ is a $k-$dimensional vector of standard normal variates and Y is distributed as $\sqrt{\chi^2/\nu}$ and $\mathbf{G}$ is a triangular matrix obtained through Cholesky's Decomposition of the variance-covariance matrix $\mathbf{V}$, such that $\mathbf{V} = \mathbf{G}\mathbf{G}'$. To obtain W_α, it is required to store all the m iterations of the random variable W, and arrange them in an ascending order. Then by using Lemma 1, we have $W_\alpha = W_{(r)}$, the upper α percentile point i.e. $P(W > W_\alpha) = \alpha$.

Now the major concern that was put forward was of W_α being a random variable, had a certain amount of variability to its solution. In other words, if it was required to run the experiment again, we would obtain a different critical point. To address this concern, Edward and Berry (1987), provided another Lemma. Let G denote the distributional function of the pivotal quantity W, and $\overline{G} = 1 - G$. Also, let $\overline{G}(W_{(r)})$ be the distribution of the conditional error probability. Then, Lemma 2 states the following

Lemma 2. Define $W_{(r)}$ as in Lemma 1. The distribution of the conditional error probability $U = \overline{G}(W_{(r)})$ over repeated simulations is the beta distribution with shape parameters $m - r + 1$ and r. That is, U has a density function given by

$$\begin{cases} \frac{\Gamma(m+1)}{\Gamma(m-r+1)\Gamma(r)} u^{m-r}(1-u)^{r-1}, & 0 < u < 1 \\ 0, & \text{elsewhere} \end{cases}$$

$\mathbb{E}(U) = \alpha$, $\mathbb{V}(U) = \alpha(1-\alpha)/(m+2)$ and for large m, this distribution is essentially normal.

Using Lemma 2, Edwards and Berry (1987) explained that with $\alpha = .05$, and simulation sizes $m + 1 = 3200,\ 80,000,\ 320,000$, the conditional coverage probability of simulation-based intervals will be $.95 \pm .01$, $.95 \pm .002$ and $.95 \pm .001$ in 99% of repeated generations respectively. Therefore, for a larger simulation size m, the variability in concern is almost negligible. To further demonstrate the benefits of the simulation based critical points, Edwards and Berry (1987) provided an efficiency study for a One-Way Analysis of Variance model and an Analysis of Covariance model.

3.1.1 Efficiency Study of a One-Way Layout

In the case of a One-Way Analysis problem, the fixed effects model is given by $Y_{ij} = \mu_i + \epsilon_{ij}$, for $i = 1, 2, \ldots, k; j = 1, 2, \ldots, n_i$ and the interval estimations of all pairwise treatment differences are $\mu_i - \mu_i'$ $(1 \leq i \neq i' \leq k)$. For such a case, the Tukey-Kramer method provides the best coverage probability for the interval estimates, in the traditional sense. Moreover, if the groups of pairwise comparisons under consideration have the same sample size, Tukey-Kramer method provides the exact probability coverage of 0.95, assuming $\alpha = 0.05$. However, in cases where

the sample sizes are not the same, Tukey-Kramer method turns out to be conservative. Hence, it was discovered that the conditional probability coverage by the simulation-based method for $m+1 = 80,000$ and $m+1 = 320,000$ was consistently closer to 0.95 when compared to Tukey-Kramer method for cases having different sample sizes. This comparison demonstrates the reliability of the simulation based method.

3.1.2 Efficiency Study of an Analysis of Covariance Layout

Now, to prove the superiority of the simulation-based method over the traditional methods in terms of sample-size savings, Edwards and Berry (1987) conducted an efficiency study for an Analysis of Covariance Model. They defined a parallel-slopes analysis of covariance setting with response Y given by $Y_{ij} = \mu_i + \gamma\chi_{ij} + \epsilon_{ij}$, for $i = 1, 2, \ldots, k; j = 1, 2, \ldots, n_i$ and the interval estimations of all pairwise treatment differences were denoted by $\mu_i - \mu_i'$ $(1 \leq i \neq i' \leq k)$. The traditional methods of multiple comparisons considered were Scheffé, Bonferroni and Šidák. To compare the performance of a simulation-based critical point with the other three methods, the concept of relative efficiency was introduced. Relative efficiency provides the approximate sample size savings at finite sample sizes by computing the ratio of the squared margins of error for the two methods under comparison, where the margin of error is the product of the critical point and the standard error. Therefore, the empirical sample size savings of method A relative to method B is $1 - (w_A/w_B)^2$, where w_A and w_B are the respective critical points. Using this concept, it was discovered that for cases with a smaller sample size, the simulation based critical point $(m+1 = 80,000)$ was 30%, 35% and 16% more efficient than Scheffé, Bonferroni and Šidák methods respectively. But, for cases having a large enough sample size, the simulation based critical point $(m+1 = 80,000)$ was 27%, 6% and 6% more efficient than Scheffé, Bonferroni and Šidák methods respectively. Therefore, Edwards and Berry (1987) claimed that the simulation-based method provided a substantial sample size savings when compared to other methods. They also mentioned that, even though the percentage savings over Bonferroni and Sidak seemed to decrease with increasing ν, for $\nu = \sum_{i=1}^{k} n_i - k - 1$, it had as an asymptote of a positive value; hence, the total savings were to increase without bound as $\nu \to 0$.

The next section discusses the extension of this idea to general response surface designs.

3.2 Simulation-Based Methods for Response Surface Designs

Sa and Edwards (1993) provided the simultaneous confidence intervals for a general response surface for $\delta(\mathbf{x})$ i.e. Expected Improvement in Mean Response w.r.t a reference blend, where

$$\delta(\mathbf{x}) = \sum_{i}^{k} \beta_i x_i + \sum_{i}^{k} \beta_{ii} x_i^2 + \sum_{i<j}^{k} \sum_{j}^{k} \beta_{ij} x_i x_j$$

for all $\mathbf{x}$ for all $\mathbf{x}$ such that $\mathbf{x}'\mathbf{x} = \sum_{i=1}^{k} x_i^2 \leq R_I^2$. Moreover, an exact solution for the critical points was given by them for $k = 1$, while they utilized an adaptation of the Scheffé simultaneous confidence intervals for $k \geq 2$. Hence, the simultaneous confidence intervals for the expected improvement in response, using the Scheffé

adaptation was given by

$$\delta(\mathbf{x}) \in \hat{\delta}(\mathbf{x}) \pm d_\alpha \mathbf{s}\{\hat{\delta}(\mathbf{x})\} \text{ for all } x_i \text{ such that } \mathbf{x}'\mathbf{x} \leq R_I^2 \tag{3.5}$$

such that the Scheffé critical point is $d_\alpha = \sqrt{(p-1)F_{\alpha,(p-1),\nu}}$, where ν is the degrees of freedom for error and p is the number of model parameters and $F_{\alpha,(p-1),\nu}$ is the upper $100\alpha\%$ critical point from the F-distribution. Sa and Edwards (1993) provided a slightly smaller simultaneous critical point by using a result from Casella and Strawderman (1980). For a 2^{nd}-Order Rotatable Design, Parody and Edwards (2007a) provided a simulation-based critical point using the methodology formulated by Edwards and Berry (1987) in the case of multiple comparisons. Hence, in equation (3.5), the critical point obtained by the Scheffé adaptation (d_α) was replaced by the simulation based critical point Q_α. This method was applied under the condition that the variance-covariance matrix $\mathbf{V}$ was a block diagonal matrix of the form

$$\mathbf{V} = \begin{bmatrix} \mathbf{V_L} & 0 & 0 \\ 0 & \mathbf{V_Q} & 0 \\ 0 & 0 & \mathbf{V_{CP}} \end{bmatrix} \tag{3.6}$$

where $\mathbf{V_L} = a\mathbf{I}_k, \mathbf{V_Q} = b\mathbf{I}_k + c\mathbf{J}_k, \mathbf{V_{CP}} = 2b\mathbf{I}_{k(k-1)/2}$ for constants a, b, c; $\mathbf{I}_k, \mathbf{J}_k$ being $k \times k$ identity matrix and $k \times k$ matrix of ones respectively. If this structure was obtainable, just like in the case of a 2^{nd}-order Rotatable Design, (3.5) could be replaced by

$$\delta(\mathbf{x}) \in \hat{\delta}(\mathbf{x}) \pm c_\alpha \mathbf{s}\{\hat{\delta}(\mathbf{x})\} \tag{3.7}$$

where $c_\alpha < \sqrt{(p-1)F_{\alpha,(p-1),\nu}}$ is the Casella and Strawderman critical point.

Parody and Edwards (2007a) improved upon the work from Sa and Edwards (1993) by using a simulation-based critical point when the design was rotatable. They defined a natural pivotal quantity for constructing $(1-\alpha)100\%$ simultaneous confidence bounds for $\delta(\mathbf{x})$ for all $\mathbf{x}$ within a specified distance R_I of the origin by

$$Q = \max_{0 \leq R \leq R_I} \max_{\mathbf{x}'\mathbf{x}=R^2} \left\{ \frac{|\widehat{\delta}(\mathbf{x}) - \delta(\mathbf{x})|}{\mathbf{s}\{\widehat{\delta}(\mathbf{x})\}} \right\} \tag{3.8}$$

They were able to obtain a form for the numerator of Q given by $(\widehat{\delta}(\mathbf{x}) - \delta(\mathbf{x}))/\sigma$, equal in distribution to

$$\sqrt{c}Z_{00}R^2 + \sqrt{a}\sum_{i=1}^{k} Z_i x_i + \sqrt{b}\sum_{i=1}^{k} Z_{ii}x_i^2 + \sqrt{2b}\sum_{i=1}^{k} Z_{ij}x_i x_j \tag{3.9}$$

where Z_i, Z_{ij}, Z_{ii} and Z_{00} were defined as mutually independent standard normal random variables.

Parody and Edwards (2007a) claimed that the advantage of utilizing a 2^{nd}-order rotatable design was that the standard error of the estimate $\widehat{\delta}(\mathbf{x})$ was constant on spheres. Moreover, for $\mathbf{V}$ of the form (3.6), $\mathbf{s}\{\widehat{\delta}(\mathbf{x})\}/\sigma$ was equal in distribution to

$$U^*(R) = \sqrt{(U/\nu)[aR^2 + (b+c)R^4]} \tag{3.10}$$

where $U \sim \chi^2_v$ independent of the Z's and $R^2 = \mathbf{x}'\mathbf{x}$

By plugging in (3.9) and (3.10) in (3.8), the form of Q was given by

$$Q = \max_{0 \le R \le R_I} \frac{1}{U^*(R)} \max_{\mathbf{x}'\mathbf{x}=R^2} \left\{ \left| \sqrt{c} Z_{00} R^2 + \sqrt{a} \sum_{i=1}^{k} Z_i x_i + \sqrt{b} \sum_{i=1}^{k} Z_{ii} x_i^2 + \sqrt{2b} \sum_{i=1}^{k} Z_{ij} x_i x_j \right| \right\}$$

The maximization of the numerator over all $\mathbf{x}$ such that $\mathbf{x}'\mathbf{x} = R^2$ was solved by Parody and Edwards (2007a), using the classic ridge analysis problem by Hoerl (1959). The critical point Q_α was obtained from Lemma 1, as mentioned in the previous section and the $(1-\alpha) \times 100\%$ simulation-based confident intervals for the estimated long term improvement were obtainable as

$$\delta(\mathbf{x}) \in \hat{\delta}(\mathbf{x}) \pm Q_\alpha \mathbf{s}\{\hat{\delta}(\mathbf{x})\} \text{ for all } x_i \text{ such that } \mathbf{x}'\mathbf{x} \le R_I^2 \tag{3.11}$$

It was also mentioned that to determine the simulation-based critical point for one-sided bounds, there wasn't a need to take the absolute value of the numerator of Q, making the computation relatively faster. Hence, for a one-sided bound, the simulation-based critical point was computed as

$$Q = \max_{0 \le R \le R_I} \max_{\mathbf{x}'\mathbf{x}=R^2} \left\{ \frac{\widehat{\delta}(\mathbf{x}) - \delta(\mathbf{x})}{\mathbf{s}\{\widehat{\delta}(\mathbf{x})\}} \right\} \tag{3.12}$$

Parody and Edwards (2007a) further reported their findings in an efficiency table. It was found out that the simulation-based method was 30% (approx.) more efficient than the Scheffé adaptation at $k = 2$. While, it was noticed that the sample size savings further increased to 110% (approx.) at $k = 5$.

The drawback of this technique was that it had some issues when dealing with regions of interest that were non-spherical in nature or if the model chosen was not a second-order model. Parody and Autin (2013) later developed a technique to optimize the amount of improvement in the long-run mean response over a reference blend based on concentric simplexes through the use of pseudocomponents. This technique would be discussed in greater detail in the following section.

3.3 Confident Visualization Techniques in the analysis of Mixture Experiments

As discussed in the previous section, Parody and Edwards (2007a) introduced a simulation based critical point for a 2^{nd}-order rotatable design. The major drawback of the technique was that their inference on $\delta(\mathbf{x})$ (expected improvement in mean response w.r.t a reference blend) couldn't directly be applied to mixture experiments. While, performing inference on the rotatable response surface designs, the reference blend was set to the point at the origin $(0, 0, \ldots, 0)$. In the case of mixture experiments, such a point doesn't exist as all the component blend proportions must add to up to unity, as per equation (2.3). Hence, Parody and Autin (2013) introduced a new technique for the creation and visualization of confidence bounds for the amount of improvement over a reference blend throughout the experimental region for the results from a mixture experiment (especially for situations in which q

> 3). The reference blend in this case, wasn't needed to be prespecified. The visualization technique involved plotting the amount of improvement versus the range of the pseudocomponents applied. It also worked for any experimental region and model of choice.

The idea behind this technique was to optimise the amount of improvement in the long-run mean response based on concentric simplexes through the use of pseudocomponents. These same ideas could be used to assess the impact of the reference blend. Parody and Edwards (2007b) discussed confident visualization techniques for high dimensional response surfaces in great detail. They demonstrated a method for visualizing the improvement contours $\delta(\mathbf{x})$ and simultaneous confidence bounds when $k \geq 2$ using canonical and ridge analysis, with examples. This technique added much needed confidence to the identification and interpretation of ridge systems. The canonical bounds allowed for the identification of flexibility in the choice of predictor values, whereas the ridge trace bounds allowed for the identification of the optimal choice of predictor values inside the experimental region. As this method wasn't prohibitive in terms of the type of design, number of predictors, radius of inference, presence of blocks and covariates, or the form of the response surface, Parody and Autin (2013) extended this technique to the domain of mixture experiments.

By applying the ridge analysis bounds, as defined by Parody and Edwards (2007b), the confidence bands for $\delta(\mathbf{x})$ along the optimal ridge path were given by $\widehat{\delta}(\hat{\mathbf{x}}_s(R)) \pm d_\alpha\{\mathbf{s}(\widehat{\delta}(\hat{\mathbf{x}}_s(R)))\}$, $0 \leq R \leq R_I$, where $d_\alpha = \sqrt{(p-1)F_{\alpha,(p-1),\nu}}$ is the scheffé adaptation of the critical point. After defining $\mathbf{x}_R$ as a $p \times 1$ vector of reference blend and $\boldsymbol{\beta}$ as a $p \times 1$ vector of parameter values, they defined the amount of improvement over the reference blend as,

$$\delta(\mathbf{x} - \mathbf{x_R}) = (\mathbf{x} - \mathbf{x_R})'\boldsymbol{\beta} \tag{3.13}$$

then, by using equation (3.13), they obtained the standard error of the estimate of $\delta(\mathbf{x} - \mathbf{x_R})$ as

$$s\left\{\widehat{\delta}(\mathbf{x} - \mathbf{x_R})\right\} = \sqrt{\hat{\sigma}^2(\mathbf{x} - \mathbf{x_R})'(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{x} - \mathbf{x_R})} \tag{3.14}$$

Now, to extract the $(1-\alpha) \times 100\%$ simultaneous confidence bands for the maximum improvement in response, the experimental region was attributed as T_I, that could be subset into l smaller regions of the same shape, denoted as T_l. This helped in providing the following expression

$$\max_{x \in T_l} \delta(\mathbf{x} - \mathbf{x_R}) \in \max_{x \in T_l} \widehat{\delta}(\mathbf{x} - \mathbf{x_R}) \pm d_\alpha s\left\{\widehat{\delta}(\mathbf{x} - \mathbf{x_R})\right\} \tag{3.15}$$

where $d_\alpha = \sqrt{pF_{\alpha,p,\nu}}$ with $F_{\alpha,p,\nu}$ being the upper $100\alpha\%$ critical point from the F-distribution with p and ν degrees of freedom.

Plots of (3.15) for the maximisation across each T_l and their respective component values against the constraint range were used to determine optimal settings. The constraint range were determined as

$$r_D = \sum_{i=1}^{q}(U_i - L_i) \tag{3.16}$$

Parody and Autin (2014) enable us to extract a large number of points inside a simplex, using concentric pseudocomponents inside the simplex region. This idea aids us to optimize the simulation-based pivotal quantity, to obtain the desired critical point.

The drawback of this research was that the confidence bounds and visualization plots were constructed using the Scheffé adaptation of the critical point, which resulted in very conservative confidence bands. To improve this technique further, we would replace the Scheffé critical point with the simulation-based critical point for obtaining tighter intervals. The visualization technique with the simulation-based critical points would be demonstrated using examples of mixture designs in Chapter 5. The next chapter introduces the research method for computing the simulation-based critical point for a $\{q, 2\}$ Simplex-Lattice design.

Chapter 4

Research Method

Monte Carlo Simulation methods to generate critical points and construction of simultaneous confidence bands are evergreen topics of discussion in the field of statistics. Over the past few decades, works of Foutz (1981); Edwards and Berry (1987); Westfall and Young (1993); Hsu (1996);and Liu, Jamshidian, and Zhang (2004) have made immense contributions towards these topics. Most recently, Han, Liu, Bretz and Wan (2015) contributed towards computing critical points to construct exact symmetric bands for a percentile line using simulation procedures. Also, Zhoua, Zhu and Wang (2018) worked on adopting a simulation based method to construct confidence bands for a percentile hyper-plane having restricted covariates. In this section we proceed to the research method being employed to generate a critical point for a $\{q, 2\}$ Simplex-Lattice Design.

4.1 Theory behind the Simulation-Based Method

Parody and Edwards (2007a) discussed the use of a natural pivotal quantity for constructing $(1-\alpha)\times 100\%$ simultaneous confidence bounds for $\delta(\mathbf{x})$ for a 2^{nd}-order rotatable response surface. By utilizing the works of Edwards and Berry (1987), Parody and Edwards (2007a) and, Parody and Autin (2013) , we extend the idea of constructing 100% simultaneous confidence bounds for the predicted response in a $\{q, 2\}$ simplex-lattice design.

Let $Y(\mathbf{x})$ be the predicted response for given observations in $\mathbf{x}$ such that $\mathbf{x}$ belongs to a particular subset or L-Pseudocomponent (T_l) inside the full simplex region. Where $T_l \in T$ such that, T is the set of all possible subsets considered in the full simplex space. The pivotal quantity Q is given by:-

$$Q = \max_{T_l \in T} \max_{\mathbf{x} \in T_l} \left\{ \frac{|\widehat{Y}(\mathbf{x}) - Y(\mathbf{x})|}{\mathbf{s}\{\widehat{Y}(\mathbf{x})\}} \right\} \tag{4.1}$$

The exact $(1-\alpha)\times 100\%$ simultaneous confidence bounds for $Y(\mathbf{x})$ is given by

$$Y(\mathbf{x}) \in \widehat{Y}(\mathbf{x}) \pm q_\alpha \mathbf{s}\{\widehat{Y}(\mathbf{x})\} \text{ for all } x_i \text{ such that } \mathbf{x} \in T_l \tag{4.2}$$

According to Edwards and Berry (1987), it is not possible to obtain a closed form solution to q_α. Hence, a random variable Q_α, generated by simulation techniques, will replace q_α. This would result in the confidence bounds having exact simultaneous coverage probability $(1-\alpha)$. Also, if the random variable Q is simulated independently m times, and if $Q_{(1)} \leq Q_{(2)} \leq ... \leq Q_{(m)}$ are the order statistics of the simulated values, then $Q_\alpha = Q_{(a)}$ will achieve this as long as α and m are chosen so that $a = (1-\alpha)(m+1)$ is an integer.

In order to proceed further, we first define the form of $Y(\mathbf{x})$ given by

$$Y(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta} + \mathbf{x}'\mathbf{B}\mathbf{x} \tag{4.3}$$

where $\mathbf{x}' = [x_1, x_2, \ldots, x_q]$, $\boldsymbol{\beta}' = [\beta_1, \beta_2, \ldots, \beta_q]$ and

$$\mathbf{B} = \begin{bmatrix} 0 & \frac{\beta_{12}}{2} & \cdots & \frac{\beta_{1q}}{2} \\ & 0 & \cdots & \frac{\beta_{2q}}{2} \\ & & \ddots & \vdots \\ (symm.) & & & 0 \end{bmatrix}$$

To simplify (8), we will require further notation. Let $\mathbf{z}' = [x_1, \ldots, x_q, x_1x_2, \ldots, x_{q-1}x_q]$ and $\boldsymbol{\gamma}' = [\beta_1, \ldots, \beta_q, \beta_{12}, \ldots, \beta_{(q-1)q}]$. Now, we can express (8) as a linear combination given by

$$Y(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta} + \mathbf{x}'\mathbf{B}\mathbf{x} = \mathbf{z}'\boldsymbol{\gamma} \tag{4.4}$$

In order to define the form of the numerator and the denominator of Q, we need to discuss the variance-covariance matrix $(\mathbf{V}\sigma^2)$ of the parameter estimates $\hat{\beta}_i$ and $\hat{\beta}_{ij}$ for all $(i, j) = 1, 2, \ldots, q; i < j$, where,

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_L & \mathbf{COV}_{L,CP} \\ \mathbf{COV}_{L,CP} & \mathbf{V}_{CP} \end{bmatrix} \tag{4.5}$$

For a $\{q, 2\}$ Simplex-Lattice Design, $Var(\hat{\beta}_i) = \sigma^2/r, Cov(\hat{\beta}_i, \hat{\beta}_{ij}) = -2\sigma^2/r, Cov(\hat{\beta}_k, \hat{\beta}_i) = Cov(\hat{\beta}_k, \hat{\beta}_{ij}) = 0$, $Cov(\hat{\beta}_{ik}, \hat{\beta}_{ij}) = 4\sigma^2/r$ and $Var(\hat{\beta}_{ij}) = 24\sigma^2/r$, r being the number of replications.

The partitioned matrices $\mathbf{V}_L$, $\mathbf{COV}_{L,CP}$ and $\mathbf{V}_{CP}$ are the variance-covariance matrices of linear coefficients; combination of linear and cross-product coefficients; and cross-product coefficients respectively. These matrices have elements in the form of coefficients of σ^2 of the above mentioned quantities. Using (9), we discuss the simulation of Q where the numerator can be defined as

$$\widehat{Y}(\mathbf{x}) - Y(\mathbf{x}) = \mathbf{x}'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \mathbf{x}'(\hat{\mathbf{B}} - \mathbf{B})\mathbf{x} = \mathbf{z}'(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \tag{4.6}$$

Using (9) and (10), we can define $\hat{\boldsymbol{\gamma}}$ as a least square estimator of the parameter vector $\boldsymbol{\gamma}$. Hence we can obtain the distribution of the estimator as $\hat{\boldsymbol{\gamma}} \sim N(\boldsymbol{\gamma}, \mathbf{V}\sigma^2)$. Following Edwards and Berry (1987), it is further noted that the signed σ-scaled numerator of Q, $(\widehat{Y}(\mathbf{x}) - Y(\mathbf{x}))/\sigma$, is equal in distribution to $\mathbf{z}'\mathbf{G}\mathbf{Z}$, where $\mathbf{G}$ is a lower-triangular matrix obtained by the Cholesky's Decomposition of $\mathbf{V}$ given by $\mathbf{V} = \mathbf{G}\mathbf{G}'$, $\mathbf{Z}$ is a standard normal vector such that $\mathbf{Z} \sim N(0, 1)$. Hence, the scalar form of the numerator is given by

$$\frac{1}{\sqrt{r}}\left\{ \sum_{i=1}^{q} a_i Z_i + \sum\sum_{i<j}^{q} a_{ij} Z_{ij} \right\} \tag{4.7}$$

where $a_i = x_i(2x_i - 1)$, $a_{ij} = 4x_ix_j$ are fixed coefficients that are only dependent on the elements of $\mathbf{x}$ and are free of error. While, Z_i, Z_{ij} are mutually independent standard normal random variables.

Moving on to the denominator of Q, we can easily ascertain that the least square estimate $\hat{\gamma}$ has a normal distribution with mean γ and variance $\mathbf{V}\sigma^2$. Moreover, $\hat{\sigma^2}$ is not dependent on $\hat{\gamma}$. Using this property we also know that $\nu\hat{\sigma}^2/\sigma^2 \sim \chi^2_\nu$. Also by (9) we are able to define the denominator of Q as,

$$\begin{aligned}\mathbf{s}\{\widehat{Y}(\mathbf{x})\} &= \sqrt{\hat{\sigma}^2\mathbf{z}'\mathbf{V}\mathbf{z}} \\ &= \sqrt{(U/\nu)\mathbf{z}'\mathbf{V}\mathbf{z}} \end{aligned} \tag{4.8}$$

Specifically, for $\mathbf{V}$ of the form (10), $\mathbf{s}\{\widehat{Y}(\mathbf{x})\}/\sigma$ is equal in distribution to

$$U^* = \sqrt{\frac{(U/\nu)}{r}\left[\sum_{i=1}^{q} a_i^2 + \sum\sum_{i<j}^{q} a_{ij}^2\right]} \tag{4.9}$$

where $U \sim \chi^2_\nu$ independent of $\mathbf{Z}$.

Hence by using the scalar form of the numerator and from (6) and (13) it follows that for a two-sided case, Q is equal in distribution to

$$Q = \max_{T_l \in T} \max_{\mathbf{x} \in T_l} \frac{|\mathbf{z}'\mathbf{G}\mathbf{Z}|}{\sqrt{(U/\nu)\mathbf{z}'\mathbf{V}\mathbf{z}}} \tag{4.10}$$

$$= \max_{T_l \in T} \max_{\mathbf{x} \in T_l} \frac{1}{U^*} \left| \frac{1}{\sqrt{r}} \left\{ \sum_{i=1}^{q} a_i Z_i + \sum\sum_{i<j}^{q} a_{ij} Z_{ij} \right\} \right| \tag{4.11}$$

$$= \max_{T_l \in T} \max_{\mathbf{x} \in T_l} \left\{ \frac{\left| \left\{ \sum_{i=1}^{q} a_i Z_i + \sum\sum_{i<j}^{q} a_{ij} Z_{ij} \right\} \right|}{\sqrt{\frac{U}{\nu}\left[\sum_{i=1}^{q} a_i^2 + \sum\sum_{i<j}^{q} a_{ij}^2\right]}} \right\} \tag{4.12}$$

According to Parody and Edwards (2007a), taking the absolute value of the numerator is not required in the case of one-sided bounds. Hence, the pivotal quantity (6) becomes

$$Q = \max_{T_l \in T} \max_{\mathbf{x} \in T_l} \left\{ \frac{\widehat{Y}(\mathbf{x}) - Y(\mathbf{x})}{\mathbf{s}\{\widehat{Y}(\mathbf{x})\}} \right\} \tag{4.13}$$

A function in R for constructing confidence intervals and one-sided bounds is given in the appendix.

Edwards and Berry (1987) addressed the concern regarding the simulation-based critical point being a random variable. It was stated that the conditional coverage probability of simulation-based confidence intervals is 0.95 ± 0.002 in 99% of the generations when the simulation size is given by $m + 1 = 80,000$. Moreover, we would realize that by using the simulation-based method instead of the other conservative methods there is a noticeable improvement of precision over existing methods by having a considerable amount of sample size savings. The benefits we reap out of the simulation-based method overshadow the concern of randomness of the critical point.

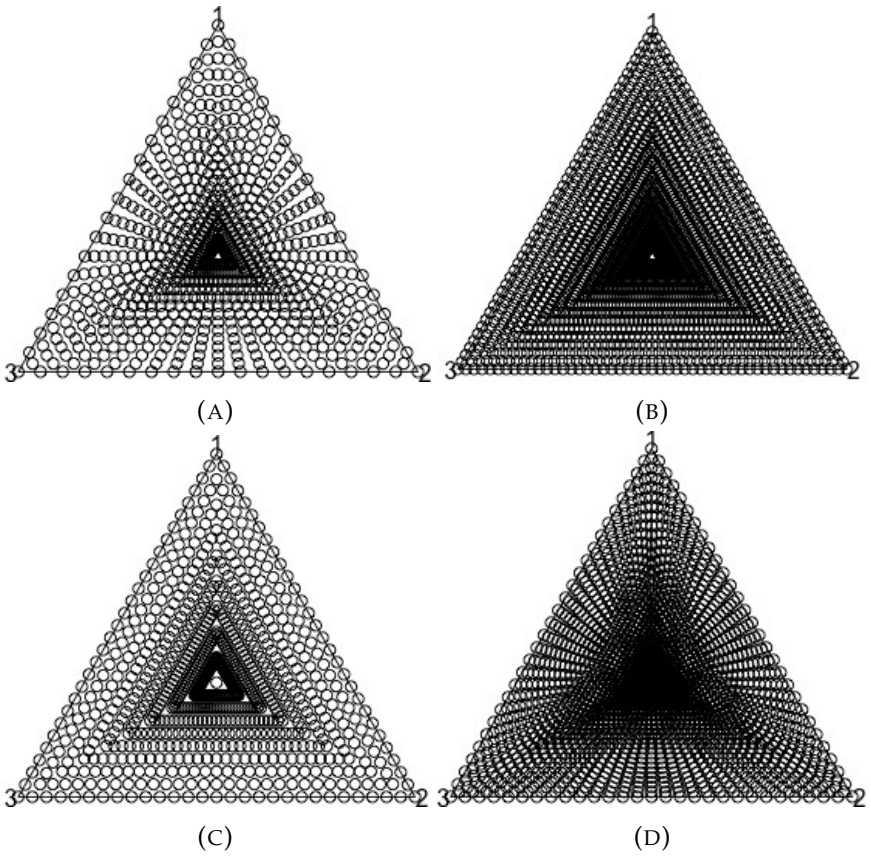
The next section talks about the L-pseudocomponents technique being utilized for the simulation procedure.

4.2 Use of L-pseudocomponents

For this experiment, the use of L-pseudocomponents was essential to apply the simulation method on a simplex-lattice design. To obtain the critical points based on the simulation procedure, it was required to optimize the pivotal quantity with respect to all possible points inside the simplex space. The idea of utilising concentric triangles for ternary mixture systems was first discussed by Cornell and Khuri (1979). They used this idea for obtaining constant prediction variance for ternary mixture systems. Moreover, this idea was generalized by Hoerl (1987) to higher dimensions for the purpose of applying ridge analysis on hypersimplexes instead of hyperspheres. Goldfarb (2004a, 2004b), Piepel and Anderson (1992) , and Piepel et al. (1993a) provided variance dispersion graphs for mixture experiments, using concentric simplexes. Piepel et al. (1993b) also used concentric simplexes to analyse response surfaces having irregularly-shaped experimental regions. Guanghui Li and Chongqi Zhang (2017) discussed a method to apply the pseudocomponent transformation on a set of uniform points under various settings of an optimal design. Guanghui Li and Chongqi Zhang (2018) adapted the random search algorithm to find optimal designs for mixture models having complex constraints. Borkowski and Piepel (2009) proposed number-theoretic methods to obtain space-filling uniform designs for high dimensional and multi-constrained mixture experiments. Lawson and Willden (2016) provided an R package to illustrate and visualize mixture designs having extreme vertices and edge and face centroids in mixture regions constrained by pseudo components. Parody and Autin (2013) favored the pseudocomponent approach to creating the points on the edge of the concentric simplexes, since the idea of pseudocomponents is well known in the mixture experiment realm.

Figure 4.1 demonstrates the effect of the number of points and L-pseudocomponents inside a simplex. In sub figures (A) and (B), keeping the number of pseudocomponents constant, the number of points on the triangle were increased. While, in sub figures (C) and (D) keeping the number of points constant, the number of pseudocomponents were increased. It is observed that the larger the number of points and L-pseudocomponents, the greater was the density of the simplex, having a better coverage.

The next chapter is based on applying the simulation based method on two different examples.



(A) (B)

(C) (D)

FIGURE 4.1: Pseudocomponents and Point coverage in the Simplex Region (l and f are indices approximating the number of points and pseudocomponents respectively)

Chapter 5

Data Examples

5.1 Artificial Sweetener Experiment

Cornell, J.A (2002) illustrated a three-component experiment consisting of three sweeteners that were glycerine, saccharin and an enhancer. The objective of the study was to determine if the possible blends of the sweeteners could be used in a popular athletic-sports drink. The amount of sweetener was fixed at 4% of the total volume (250 mL.) of the sports drink.

In Table 1, the 3 sweeteners are given as the 3 components x_1, x_2 and x_3, where the values associated with them are the design points considered on the simplex region. The response y represents the "intensity of sweetness aftertaste" score for each blend. This was measured as a score on the scale of 1-30. The score of 1 being "no aftertaste" and 30 being " very extreme aftertaste". The values associated with the response y were computed by averaging out the scores of 20 respondents in a survey.
By fitting model (2) to the 15 data values at the six blends (1-6) of Table 1., the parameter estimates are given by

$$\hat{\beta}' = [10.40, 6.15, 3.90]$$

$$\hat{\mathbf{B}} = \begin{bmatrix} 0 & 13.385 & 10.035 \\ 13.385 & 0 & 14.485 \\ 10.035 & 14.485 & 0 \end{bmatrix}$$

The MSE of the fitted model is 0.3206 with 9 df.

TABLE 5.1: Data from the Artificial Sweetener Experiment

Blend	Glycine x_1	Saccharine x_2	Enhancer x_3	y
1	1	0	0	10.1, 10.7
2	0	1	0	5.8, 6.5
3	0	0	1	4.2, 3.6
4	1/2	1/2	0	14.5, 15.4, 15.0
5	1/2	0	1/2	12.9, 12.0, 11.6
6	0	1/2	1/2	11.6, 13.0, 12.2

As minimization is our goal, we are only concerned with the upper bound, therefore we will utilize a simulation-based one-sided confidence bound. For a simulation

size of $m + 1 = 80,000$ and $\alpha = 0.05$, the critical point for the simulation-based one-sided confidence bound was computed as $Q_{0.05} = 3.244$. Figure 3(a) presents the estimated improvement contours $\hat{\delta}(\mathbf{x})$, whereas Figure 3(b) shows the 95% simultaneous upper confidence bound for the amount of improvement.

To visualize estimated maximum improvement and 95% Upper Confidence Bounds, we utilize a confident visualization technique provided by Parody and Autin (2013) which enables us to observe and interpret the contours of the entire surface inside the simplex region. In Figure 3(a), we can see that there is indeed a region that yields positive values for the estimated improvement. Any of the component values inside the contour with response value 0 will meet this requirement. The improvement region is closer to the left vertex, proving a better response than the control settings (centroid). While, in Figure 3(b), we observe that the region that yields positive values of improvement is larger than that of the estimated maximum improvement. In fact, there is also a region where the 95% Upper Confidence bounds for estimated improvement are greater than 0.5. Having a minimization problem, based on figure 3, we observe that the minimum response is obtained towards the x_3 vertex. Hence, the minimum estimated response is found out to be 3.9 corresponding to the design point $(0, 0, 1)$.

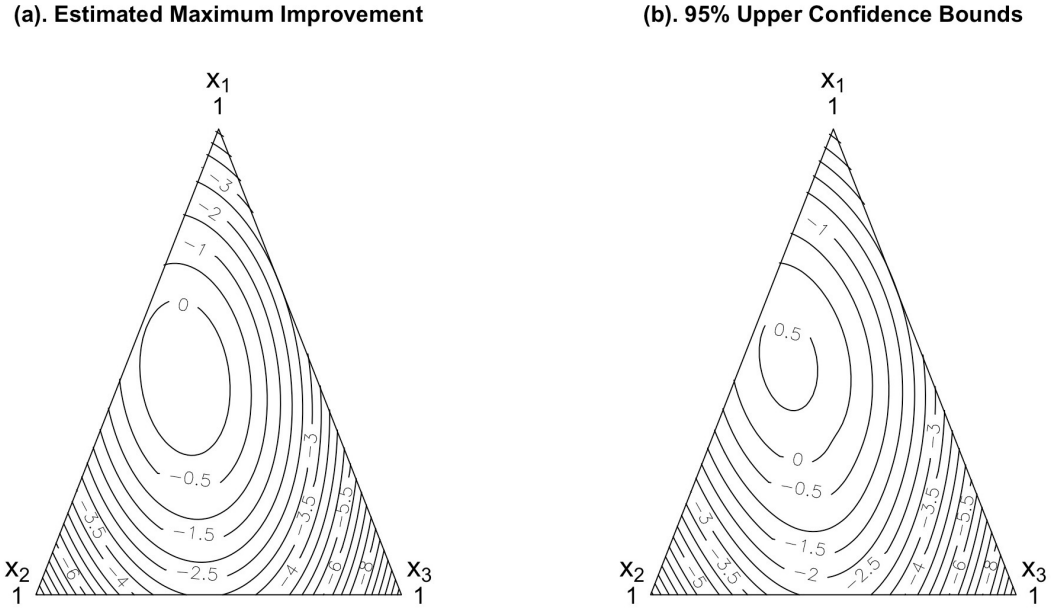


FIGURE 5.1: Artificial Sweetner Example; (a) Estimated Improvement contours relative to the centroid; (b) simulation-based lower 95% simultaneous confidence bounds. The region inside the zero contour indicates improvement over the control settings

In this example, we observe that the squared simulation-based critical point $Q^2_{0.05} = 10.524$ is approximately half in magnitude to the squared critical point obtained from the Scheffé method, $d^2_{0.05} = 20.243$. We would now use the concept of relative efficiency to demonstrate the reason of this result being desirable. Relative efficiency is computed by taking the ratio of the squared margin of errors for the methods under consideration. In this case we have equal standard errors of the estimates for all $\mathbf{x}$. Hence, the relative efficiency in this case would just be the ratio of squared critical points. To evaluate the percentage increase, we would subtract 1 from the

ratio. Hence we have $(d^2_{0.05}/Q_{0.05})^2 - 1 = 0.92$. This would mean that to make the scheffé method equally precise to the simulation-based method, it will be required to increase the sample size by 92%. 2

5.2 Tropical Beverage Experiment

The second example is also an experiment discussed by Cornell, J.A (2002). In this experiment, a tropical beverage was formulated by blending the juices of watermelon (x_1), orange (x_2), pineapple (x_3), and grapefruit (x_4). The response measured in this study is the average flavor score (based on a scale of 1-9) considering 40 samples of each blend having 3 replicates each. The goal of this study is to maximize the average flavor score of the tropical beverage. Each of the fruit flavors were considered as pure blends as well as having binary combinations with the other three flavors.

TABLE 5.2: Data from the Tropical Beverage Experiment

Blend	Watermelon x_1	Orange x_2	Pineapple x_3	Grapefruit x_4	Average Flavor Scores y
1	1	0	0	0	5.68, 5.99, 5.74
2	0	1	0	0	6.00, 5.52, 6.05
3	0	0	1	0	5.41, 6.15, 5.56
4	0	0	0	1	5.13, 4.53, 4.53
5	1/2	1/2	0	0	7.00, 6.81, 7.16
6	1/2	0	1/2	0	8.00, 7.51, 7.08
7	1/2	0	0	1/2	6.19, 5.67, 6.14
8	0	1/2	1/2	0	5.89, 5.95, 5.89
9	0	1/2	0	1/2	5.68, 5.07, 5.53
10	0	0	1/2	1/2	5.64, 5.00, 5.90

By fitting model (2) to the 10 data values at the six blends (1-10) of Table 2., the parameter estimates are given by

$$\hat{\boldsymbol{\beta}}' = [5.80, 5.85, 5.71, 4.73]$$

$$\hat{\mathbf{B}} = \begin{bmatrix} 0 & 2.32 & 3.55 & 1.46 \\ 2.32 & 0 & 0.26 & 0.27 \\ 3.55 & 0.26 & 0 & 0.59 \\ 1.46 & 0.27 & 0.59 & 0 \end{bmatrix}$$

The MSE of the fitted model is 0.1023 with 20 df.

We assume that the objective of the study is to see if any improvement in the score for the reference blend can be made over the average flavor scores. The reference blend was set as the centroid for the Tropical Beverage example i.e. $\mathbf{x}'_R = (0.25, 0.25, 0.25, 0.25)$. The estimated response for the reference blend is $\hat{y}(\mathbf{x}_R) = 6.579$. For this experiment, the simulation based critical point is given by $Q_{0.05} = 3.309$. As we have considered a full simplex design, the range for the component values is $(0, 1)$.

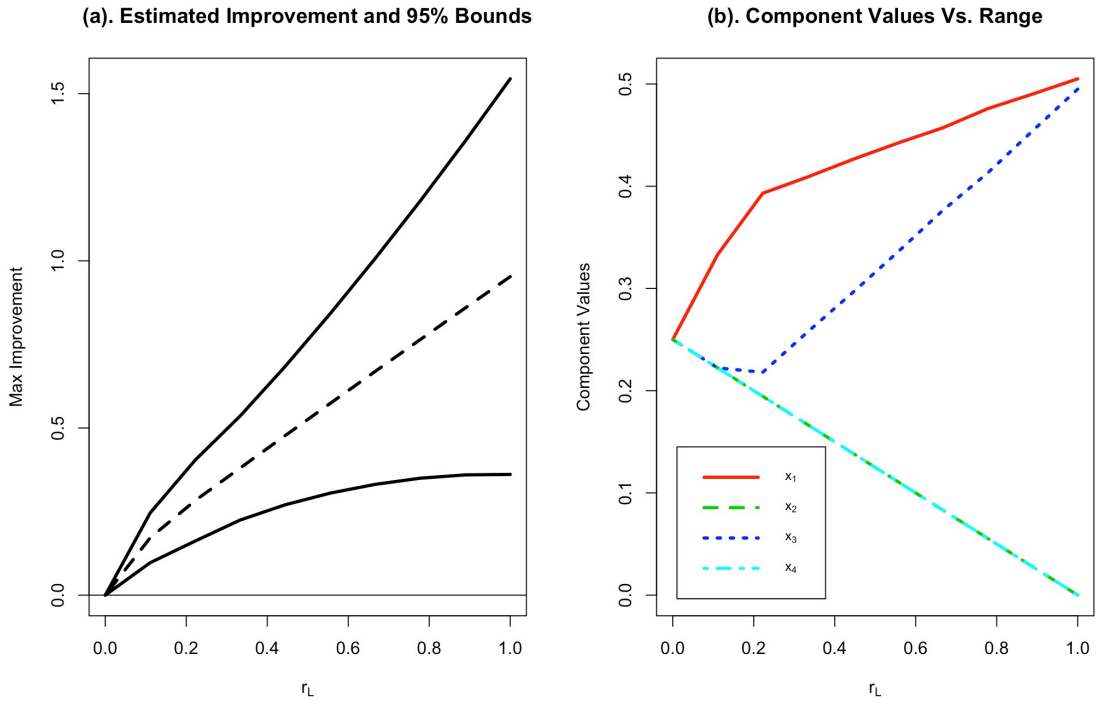


FIGURE 5.2: Tropical Beverage Example; (a) 95% simultaneous bounds for the amount of improvement over the control along the estimated optimal component path using the simulation-based method (4); (b) estimated optimal component path

Based on Figure 4(a), the estimated response for the Tropical Beverage experiment is maximized at $r_L = 1$. This corresponds to the edge of the experimental region. At this range, the estimated amount of improvement over the reference blend is roughly 1.0. Based on Figure 7(b), the blend where the maximum is found is $(0.505, 0, 0.495, 0)$. The lower bound for improvement for the top flavor score at this blend is 0.36. The fact that the entire lower bound region is made up of positive values indicates that there is indeed some possible improvement in top contour score over the reference blend.

In this example, having $q = 4$, we see more improvement in efficiency when we compare the squared simulation-based critical point $Q^2_{0.05} = 10.95$ with the Scheffé adaptation given by Sa and Edwards (1993), $d^2_{0.05} = 27.737$. The relative efficiency in this case is computed as $(d^2_{0.05}/Q_{0.05})^2 - 1 = 1.53$ i.e. for the scheffé method to have the same precision compared to the simulation-based method, it would require an increase in the sample size by 153%.

Chapter 6

Discussion and Conclusions

Based on the examples provided in the previous section, the simulation-based method defined in section 3 yields substantially narrower bounds than the Sa and Edwards (1993) adaptation of the Scheffé method. In this section, an efficiency study is conducted to ascertain the amount of sample size savings by using the simulation-based method for confidence intervals. The study compared critical points for $q = 3 - 5$, $\alpha = 0.05$ and $r = 2, 3, 4, 5, 7, \infty$. Simplex lattice designs with $m = 2$ were utilized in the efficiency study, for all the values of q and r mentioned above. Table 3 provides the sample-size savings of the two-sided simulation-based method over the Sa and Edwards adaptation of the Scheffé method.

TABLE 6.1: Approximate sample-size savings, two-sided simulation-based method to the Sa and Edwards (1993) adaptation of the Scheffé method at $\alpha = 0.05$.

	q		
r	3	4	5
2	50.1%	78.8%	108.3%
3	45.9%	75.4%	97.4%
4	44.9%	72.9%	95.8%
5	44.8%	73.3%	93.9%
7	43.6%	71.0%	94.0%
∞	43.6%	69.5%	92.3%

From Table 3, we observe an improvement of more than 100% over the Sa and Edwards adaptation of the Scheffé method by using this simulation-based method when we set the number of factor to be large enough. Even for a small number of factors , the sample size savings is still considerable, at approximately 50%. As we increase the number of pure blends (q), the sample size savings have greater improvement over the Scheffé method. For $q = 5$, the sample size savings are more than double for using the simulation-based critical point.

Considering the examples in Section 4, we have a better sample size improvement for one-sided confidence bounds, compared to those constructed using the Scheffé adaptation. A substantial amount of work is yet to be done, as this paper introduces simulation based inference to the domain of mixture experiments considering the simple case of a $\{q, 2\}$ simplex-lattice design having a polynomial model of degree 2. It is of further interest to work upon higher degree models and other forms of mixture experiments. The biggest challenge to achieve this goal is the complications in the denominator in (6). The ultimate objective of this study is to generalize this method across all forms of mixture experiments and response surface designs. The

authors are currently working on these research topics and look forward to further participation and comment.

Bibliography

[1] Casella, G. and Strawderman, W.E. (1980) 'Confidence bands for linear regression with restricted predictor variables',*Journal of the American Statistical Association* Vol. 75, No. 372, pp.862–868.

[2] Cornell, J.A. (2002) *Experiments with Mixtures Designs, Models, and the Analysis of Mixture Data* , 3rd ed., Wiley & Sons, New York.

[3] Cornell, J.A. and Khuri, A.I. (1979) 'Obtaining constant prediction variance on concentric triangles for ternary mixture systems', *Technometrics* , Vol. 21, No. 2, pp.147–157.

[4] Edwards, D. and Berry, J. J. (1987). The efficiency of simulation-based multiple comparisons. *Biometrics*, 43, 913-928.

[5] Foutz, R. V. (1981). Simultaneous tests for finite families of hypotheses.*Communications in Statistics: Theory and Methods*, 11, 1839-1853.

[6] Goldfarb, H.B., Borror, C.M., Montgomery, D.C. and Anderson-Cook, C.M. (2004a) 'Three-dimension variance dispersion graphs for mixture-process experiments', *Journal of Quality Technology*, Vol. 36, No. 1, pp.109–124.

[7] Goldfarb, H.B., Borror, C.M., Montgomery, D.C. and Anderson-Cook, C.M. (2004b) 'Fraction for design space plots for assessing mixture and mixture-process designs', *Journal of Quality Technology*, Vol. 36, No. 2, pp.169–179.

[8] Guanghui Li & Chongqi Zhang (2017) The pseudo component transformation design for experiment with mixture, *Statistics and Probability Letters*, Volume 131, Pages 19–24

[9] Guanghui Li & Chongqi Zhang (2018) Random search algorithm for optimal mixture experimental design, *Communications in Statistics - Theory and Methods*, 47:6, 1413-1422, DOI: 10.1080/03610926.2017.1321122

[10] Hoerl, A. E. (1959). Optimum solutions of many variable equations. *Chemical Engineering Progress*, 55, 69-78.

[11] Hsu, J. C. (1996).*Multiple Comparisons Theory and Methods.*London: Chapman & Hall.

[12] John J. Borkowski & Greg F. Piepel (2009) Uniform Designs for Highly Constrained Mixture Experiments, *Journal of Quality Technology*, 41:1, 35-47, DOI: 10.1080/00224065.2009.11917758

[13] John Lawson, Cameron Willden (2016). Mixture Experiments in R Using mixexp. *Journal of Statistical Software*, Code Snippets, 72(2), 1-20., "doi:10.18637/jss.v072.c02"

[14] Liu, W., Jamshidian, M. and Zhang, Y. (2004). Multiple comparison of several regression models.Journal of the American Statistical Association,99, 395-403.

[15] Parody, R. and Autin, M. (2013) 'Confident visualization techniques in the analysis of mixture experiments',*Int. J. Experimental Design and Process Optimisation International Journal of Statistics and Management Systems, Vol. 3, No. 3, pp.245–262.*

[16] Parody, R.J. and Edwards, D. (2007a) 'Simulation-based inference on the improvement in a rotatable response surface', *Quality Technology and Quantitative Management*, Vol. 4, No. 4, pp.489–499.

[17] Parody, R.J. and Edwards, D. (2007b) 'Confident visualization techniques for improvement in high dimensional response surfaces', *International Journal of Statistics and Management Systems,*Vol. 1, Nos. 1–2, pp.112–129.

[18] Piepel, G. and Anderson, C.M. (1992) 'Variance dispersion graphs for designs on polyhedral regions',*Proceedings of the Section on Physical and Engineering Sciences*, American Statistical Association, pp.111–117.

[19] Piepel, G., Anderson, C.M. and Redgate, P.E. (1993a) 'Variance dispersion graphs for designs on polyhedral region – revisited', Proceedings of the Section on Physical and Engineering Sciences, American Statistical Association, pp.102–107.

[20] Sa, P. and Edwards, D. (1993) 'Multiple comparisons with a control in response surface methodology', *Technometrics* , Vol. 35, No. 4, pp.436–445.

[21] Sanyu Zhoua, Jingjing Zhu, Defa Wang (2018) Simultaneous confidence bands for a percentile hyper-plane with covariates constrained in a restricted range, *Journal of Computational and Applied Mathematics*, Volume 344, Pages 301-312

[22] Westfall, P.H. and Young, S.S.(1993). *Resampling - Based Multiple Testing Examples and Methods for p-Value Adjustment.* New York: Wiley.

[23] Yang Han, Wei Liu, Frank Bretz, Fang Wan (2015) Simultaneous confidence bands for a percentile line in linear regression, *Computational Statistics & Data Analysis*, Volume 81, Pages 1-9

Appendix A

Simulation Code

```
################################
################################
################################
### Function to generate the complete design matrix using the design points
##(will be called inside the main algo)
full.mat<-function(x){
  x.binary<-NULL
  for(i in 1:ncol(x)){
    for(j in 1:ncol(x)){
      if(i<j){
        x.binary<-cbind(x.binary,x[,i]*x[,j])
      }
    }
  }
  X<-cbind(x,x.binary)
  return(X)
}
#############################
#############################
#############################
## Function to generate all possible points inside
##the simplex region (will be called inside the main algo)
ridge.mix<-function(k=3,l=10){

  theta<-seq(0,0.5,length=l)

  #browser()
  x<-expand.grid(rep(list(theta),(k-2)))
  x<-cbind((1-(apply(x,1,sum))),x)
  x.mid<-x
  for(ii in 2:(k-1)){
    x.mat<-x
    x.mat[,1]<-x[,ii]
    x.mat[,ii]<-x[,1]
    x.mid<-rbind(x.mid,x.mat)
  }

  x.mid<-unique(x.mid)

  x.mid<-cbind(matrix(0,nrow(x.mid),1),x.mid)
```

```
  #browser()

  x.final<-x.mid

  for(ii in 2:(k)){
    x.mat1<-x.mid
    x.mat1[,1]<-x.mid[,ii]
    x.mat1[,ii]<-x.mid[,1]
    x.final<-rbind(x.final,x.mat1)
  }

  #browser()
  return(unique(x.final))

}

#################################
#################################
#################################
## THE MAIN ALGORITHM (2-tailed)
##(used for confidence interval)
w.two.sim<-function(x.design,B,k=3,alpha=0.05,l=10,f=10,seed=101)
  ####{ # k=q= no. of dimensions21
  ####Intialization
  #x.design = design matrix without cross products
  #B= no. of simulations
  #k=q= defines a (q-1) dimensional simplex
  #l = index of no. of points on the simplex
  #f= index of no. of pseudocomponents inside the original simplex
  ####
  start<-proc.time()
  set.seed(seed)
  ####### Computing the complete design matrix,
  ####variance-covariance matrix and Cholesky
  X<-full.mat(x.design)
  rows<-nrow(X)
  cols<-ncol(X)
  df = nrow(X)-ncol(X)
  XX<-t(X)%*%X
  invXX<-solve(XX, diag(x=1,nrow = cols,ncol = cols))
  g.mat<-t(chol(invXX))
  ####################################################
  ################Initializing the matrices of random numbers (Z, chi-sq)
  L<-as.matrix(seq(0,(1/k),length=f))
  ran.mat.1<-matrix(rnorm(cols*B),nrow = cols,ncol = B)
  ran.mat.2<-matrix(rchisq(B,df)/df,nrow = B, ncol=1)
  x.tri<-as.matrix(ridge.mix(k=k,l=l))
  sim.max<-NULL
  ################################################
  #####The simulation loop (B=no. of simulations)
```

```
  for(j in 1:B){
    crit.temp<-0
    max.pivot.mat<-NULL
    ##### The loop that applies transformation to the points on
    ###the full simplex, to generate smaller and smaller simplexes
    ##inside the original simplex
    for(z in 1:f){
      x.mat<-NULL
      pivot.mat<-NULL
      x.mat<-as.matrix((x.tri*(1-(k*L[z])))+L[z]) # The Transformation
      x.mat<-full.mat(x.mat)
      r<-nrow(x.mat)
      ####### The loop to compute the pivotal quantity
      ###for each row of the design space
      for(i in 1:r){
        num<-abs(as.matrix(t(x.mat[i,]))%*%g.mat%*%ran.mat.1[,j])
        den<-sqrt(ran.mat.2[j,]*(as.matrix(t(x.mat[i,])))
        %*%invXX%*%as.matrix(x.mat[i,])))
        crit<-num/den
        ####### The IF condition to carry out the double maximizations
        if(crit>crit.temp){
          crit.temp<-crit
        }
        else{
          next
        }
      }
    }
    ##############################################
    sim.max<-rbind(sim.max,crit.temp)
  }
  max.sim.max<-max(sim.max)
  #########Sorting and Pulling off the alpha-percentile from the simulations
  max.sim.sort<-as.matrix(sort(sim.max))
  sim.percentile<-max.sim.sort[(1-(alpha/2))*B,1]
  ##############################################3#
  elapsed<-proc.time()-start
  return(list(max = max.sim.max, sim.percentile =sim.percentile,time=elapsed))
}
```